https://doi.org/10.64211/oidaijsd191102

Process-Informed Latent-State Discovery and Interpretable Ensemble Learning for Full-Scale Wastewater Treatment Monitoring

Nguyen-Quang Thanh [1], Nguyen-Ba Hoang [*2]
[1,2] Vung Tau University, Ho Chi Minh City 700000, Vietnam.
*Corresponding author: hoangnb@bvu.edu.vn

OIDA International Journal of Sustainable Development, Ontario International Development Agency, Canada.
ISSN 1923-6654 (print) ISSN 1923-6662 (online) www.oidaijsd.com
Deposited in SSRN repository: https://www.ssrn.com/ssrn/

Abstract: Reliable wastewater treatment monitoring requires methods that convert multivariate sensor streams into operationally meaningful states. This study proposes a process-informed, cluster-guided framework integrating unsupervised regime discovery, semantic state labeling, and interpretable tree-based classification. Online hydraulic, aeration, thermal, dissolved-oxygen, nitrate, and ammonium data from the full-scale Tilburg wastewater treatment plant were analyzed. Clustering identified four recurrent conditions: low-load operation, stable nitrification, oxygen-limited operation, and shock-load/transition conditions. The derived labels were predicted from routinely available online inputs using Decision Tree, Random Forest, and XGBoost classifiers, which achieved accuracies of 0.960, 0.988, and 0.989, respectively. Misclassifications were concentrated near boundaries between neighboring states, indicating gradual process transitions rather than widespread model instability. Model interpretation linked predictions to physically meaningful drivers, particularly influent loading, staged aeration, and temperature. The framework therefore provides an accurate, interpretable basis for real-time monitoring, early warning, operational decision support, and future integration with plant control systems in practical applications.

Keywords: latent operating regimes; semantic cluster labeling; full-scale WWTP analytics; nitrogen-conversion dynamics; aeration-control intelligence; explainable tree ensembles; weakly labeled process data

Introduction

Reliable operation of wastewater treatment plants increasingly depends on continuous monitoring and timely interpretation of process conditions before water-quality deviations propagate through the treatment train [1]. In biological treatment systems, dissolved oxygen, ammonium, and nitrate are not only effluent-quality indicators but also direct markers of nitrification, denitrification, aeration efficiency, and load adaptation [2, 3]. This study examines this problem using online sensor data from the Tilburg wastewater treatment plant in the Netherlands, a full-scale facility with parallel biological treatment lines, aeration control, and monitored variables linked to oxygen and nitrogen transformation dynamics [4].

Despite the growing availability of online monitoring systems, converting raw wastewater sensor streams into robust operational knowledge remains challenging. Wastewater process data are typically multivariate, nonlinear, and affected by interactions among influent flow, aeration intensity, temperature, and biological state [5, 6]. In the Tilburg dataset, the input space includes total influent flow, multiple aeration-flow variables, and temperature, whereas the monitored process responses include dissolved oxygen in successive aerobic tanks, nitrate, and ammonium at the end of the aerobic zone [7, 8]. Although this structure is rich in operational information, explicit state labels are not directly available, limiting the use of conventional supervised classification [9, 10]. At the same time, direct regression of individual outputs such as dissolved oxygen, ammonium, or nitrate does not fully address the practical need to

recognize whether the plant is operating under stable, transitional, oxygen-limited, or load-sensitive conditions [11–13].

This study addresses this gap by formulating wastewater operational-state prediction as a cluster-guided machine learning problem. Instead of treating the data only as a continuous input–output regression task, latent operating patterns are first identified from the multivariate process space and then converted into interpretable operational-state labels [14, 15]. These labels transform originally weakly structured online sensor data into a supervised learning dataset that can support deployable state prediction. The framework was developed using two temporal periods from the Tilburg plant: a nine-month development period for clustering, labeling, model training, validation, and internal testing, followed by a three-month evaluation period in which output variables are unavailable and plant states must be inferred from online inputs [16, 17].

Within this framework, unsupervised clustering was used to reveal recurring operating regimes, after which Decision Tree, Random Forest, and Extreme Gradient Boosting classifiers were trained to predict future operational states from online measurements. These tree-based models were selected because they combine nonlinear learning capacity, robustness for heterogeneous tabular data, and practical interpretability for engineering applications. The resulting workflow links data-driven state discovery with supervised prediction and process-oriented explanation, allowing the learned model to support not only classification accuracy but also operational understanding.

The contribution of this study is fourfold. First, it establishes a cluster-guided framework that integrates unsupervised state discovery, semantic label construction, supervised classification, and interpretability within a single workflow. Second, it demonstrates how supervised operational-state labels can be systematically derived from weakly labeled wastewater sensor data. Third, it provides a comparative evaluation of three tree-based classifiers under the same chronological data structure. Fourth, it connects model interpretation with wastewater-process mechanisms, thereby supporting the translation of high-dimensional sensor streams into operationally meaningful knowledge for real-time monitoring and decision support.

Full-scale wastewater treatment system and sensor dataset

Full-scale plant configuration and operating context

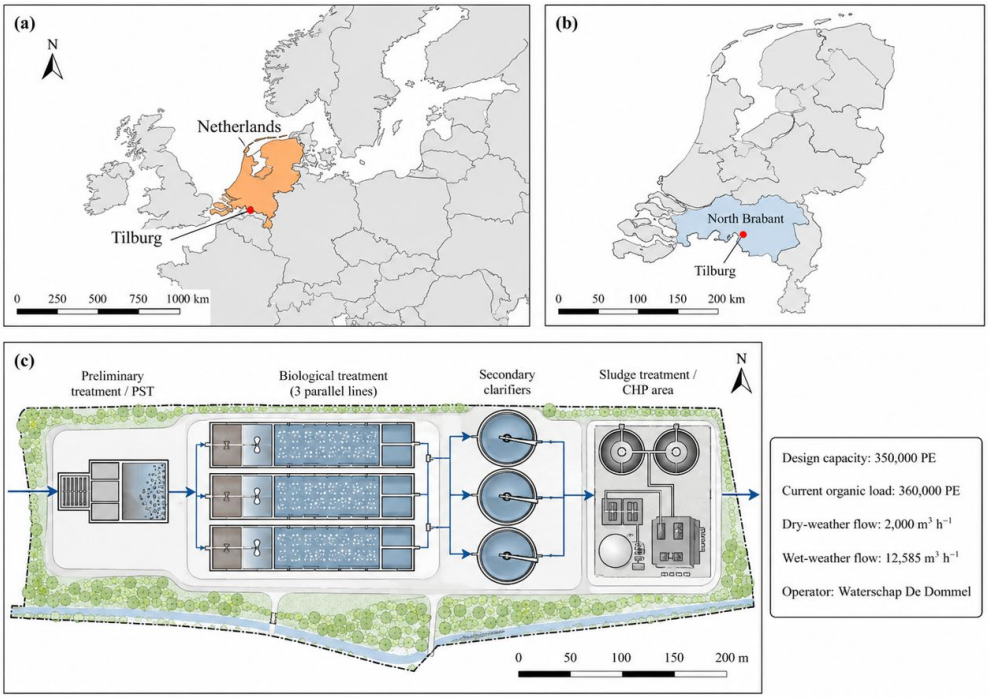


Fig. 1. Geographic setting and plant-scale configuration of the Tilburg wastewater treatment plant: (a) location within Europe, (b) location within the Netherlands, and (c) treatment-train layout and design characteristics

The case study was conducted at the Tilburg wastewater treatment plant in the southern Netherlands. As shown in **Fig. 1**, Tilburg is the second-largest wastewater treatment plant operated by Waterschap De Dommel, with a design capacity of approximately 350,000 population equivalent and a current organic loading of about 360,000 population equivalent [18, 19]. The plant treats approximately 2,000 m³/h under dry-weather conditions and up to 12,585 m³/h during wet-weather operation [20, 21], indicating substantial hydraulic variability across operating conditions. In addition to the water-treatment line, the facility includes sludge treatment and energy recovery through combined heat and power generation [11]. Its water line comprises three parallel treatment systems with primary settling, biological treatment, chemical phosphorus removal, and secondary settling [13, 16]. This full-scale configuration provides a suitable operational setting for developing a data-driven framework to identify and predict wastewater treatment states from online sensor measurements [22–24].

Biological treatment train and monitoring architecture

Wastewater entering the Tilburg plant first passes through preliminary treatment, including screening and grit removal, before being conveyed to the primary settling tank and then to the activated-sludge process [25, 26]. As illustrated in **Fig. 2**, the influent is distributed into three parallel biological treatment lines, denoted AT1, AT2, and AT3. Each line consists of an aerobic selector tank, a denitrification tank, a facultative tank, and three consecutive aerobic nitrification tanks. This configuration defines the main hydraulic and biochemical pathway through which influent loading, oxygen supply, and nitrogen conversion evolve across the treatment train [12, 27].

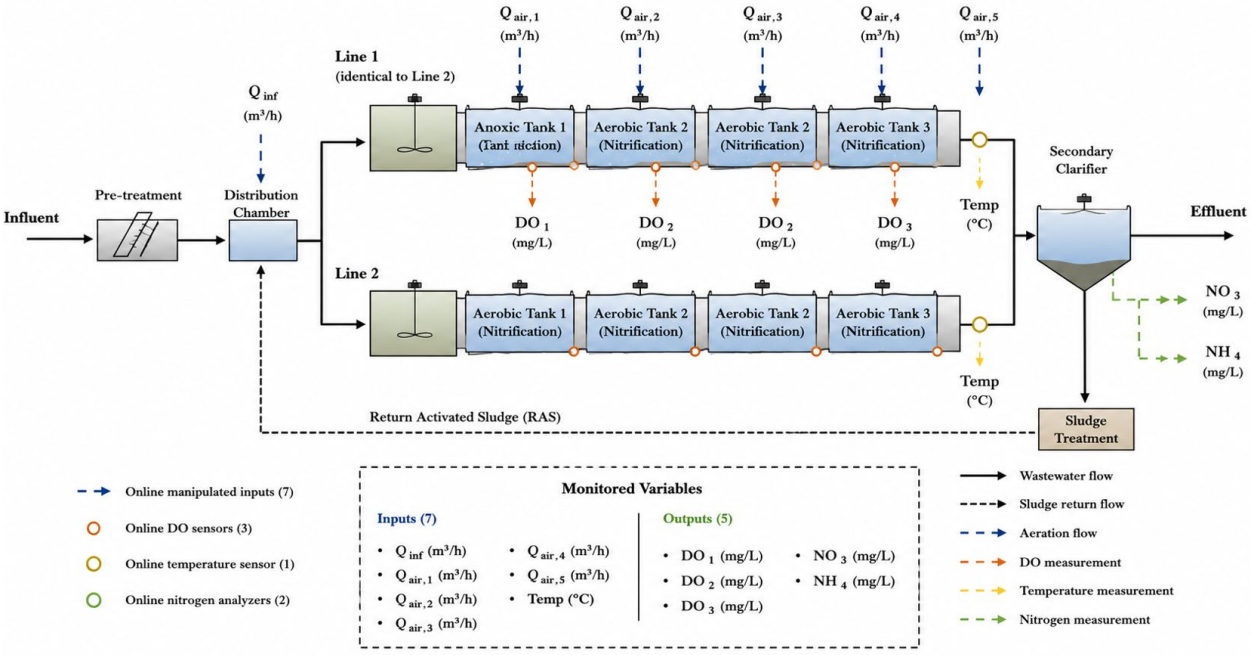


Fig. 2. Biological treatment train, online monitoring points, and modeled hydraulic, aeration, thermal, and nitrogen-conversion variables at the Tilburg wastewater treatment plant

The plant operation is further governed by internal recycle and sludge-return flows. Mixed liquor is recirculated from the end of the aerobic zone to the head of the denitrification tank, while return activated sludge is also introduced into the biological process [28, 29]. Aeration is controlled through a cascade ammonium–dissolved oxygen strategy, and the mixed liquor suspended solids concentration in the aeration tanks is maintained at approximately 4.5–5.0 g/L. The reported average internal nitrate-recycle flow, return activated-sludge flow, and waste-sludge flow are 9,800 m³/h, 1,600 m³/h, and 30.5 m³/h, respectively. The corresponding tank volumes are 10,920 m³ for the anaerobic tank, 10,100 m³ for the facultative tank, and 11,179 m³ for each of the three aeration tanks [30, 31]. These values confirm that the observed sensor dynamics originate from a coupled full-scale biological treatment system rather than from a simplified experimental setting.

Chronological dataset and deployment scenario

The study used the online/offline Tilburg wastewater treatment dataset, which is organized into two chronological phases [17, 34]. The first phase is a nine-month development period from May 2021 to January 2022, in which both

online inputs and process-response variables are available for exploratory analysis, clustering, label generation, model training, validation, and internal testing [35]. The second phase is a three-month evaluation period from March 2022 to June 2022, in which the output variables are not provided and operational states must be inferred from the available online inputs [36]. This two-phase structure supports the study objective because it separates state discovery and model construction from deployment-like state inference.

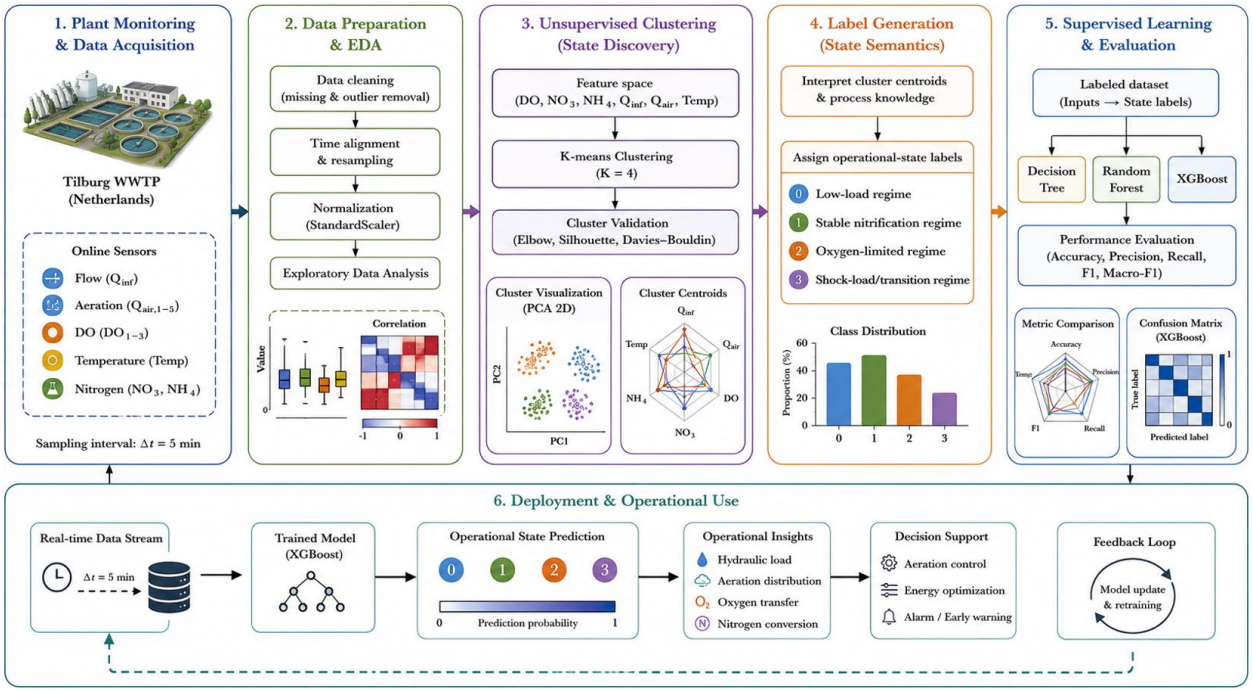


Fig. 3. *End-to-end process-informed cluster-guided learning framework for latent-regime discovery, semantic state labeling, supervised prediction, model explanation, and operational deployment*

As shown in **Fig. 3**, the dataset structure directly supports the proposed cluster-guided workflow. The development phase provides the complete historical record required to identify latent operating regimes and transform them into supervised operational-state labels, whereas the later phase represents a realistic monitoring condition in which future plant states must be predicted from online measurements alone. Thus, Fig. 3 does not only summarize the data flow, but also clarifies the methodological link between sensor acquisition, latent-state discovery, supervised learning, and online operational-state prediction.

Online inputs, process responses, and operational roles

The online input variables used in this study comprise the total influent flow rate, Q_{inf}, five staged aeration-flow variables, $Q_{air,1}$–$Q_{air,5}$, and the temperature measured at the end of the aerobic zone, denoted as Temp. Specifically, $Q_{air,1}$represents airflow to the aerobic selector tank, $Q_{air,2}$corresponds to airflow to the facultative tank, and $Q_{air,3}$, $Q_{air,4}$, and $Q_{air,5}$represent airflow supplied to the first, second, and third aerobic nitrification tanks, respectively [42, 43]. Together, these seven inputs describe the main hydraulic, aeration-control, and thermal conditions governing the biological treatment process.

The process-response variables include dissolved oxygen concentrations in the three aerobic tanks, denoted as DO_1, DO_2, and DO_3, together with nitrate concentration, NO_3, and ammonium concentration, NH_4, measured at the end of the aerobic zone [44]. These five outputs summarize oxygen-transfer status and nitrogen-conversion performance along the biological treatment line. Because dissolved oxygen and nitrogen species are directly linked to nitrification efficiency, denitrification behavior, and aeration control, they provide the biochemical basis for identifying latent operational states in the monitored wastewater treatment process.

Operational-state learning problem

The original dataset contains continuous online input variables and process-response measurements. However, this study did not formulate the task as direct numerical prediction alone. Instead, the analytical workflow first identified

latent operating groups through clustering, then converted these groups into operational-state labels for supervised learning [45, 46]. This design transformed the weakly structured wastewater sensor data into a state-based classification problem.

Given the online sensor input vector at time t,

$$\mathbf{X}_t = \left[Q_{\text{inf},t}, Q_{\text{air},1,t}, Q_{\text{air},2,t}, Q_{\text{air},3,t}, Q_{\text{air},4,t}, Q_{\text{air},5,t}, \text{Temp}_t\right], \quad (1)$$

the objective is to predict the corresponding operational-state label

$$y_t \in \{1,2,\dots,K\}, \quad (2)$$

where K denotes the number of latent operational states identified from the clustering stage. Under this formulation, the continuous response variables DO_1, DO_2, DO_3, NO_3, and NH_4were not discarded; rather, they were used to characterize and interpret the operational meaning of each state during label generation.

Accordingly, the main contribution of the study lies in converting multivariate online wastewater measurements into an interpretable supervised classification task. The continuous process responses may still support auxiliary regression or post hoc validation, but the primary objective is operational-state prediction in a form that is directly aligned with real-time monitoring and decision support [48–50].

rocess-informed cluster-guided state-learning framework

Framework overview and problem formulation

The proposed methodology converted multivariate online wastewater measurements into an interpretable state-prediction framework through a sequential workflow comprising preprocessing, exploratory analysis, unsupervised state discovery, cluster-guided label generation, supervised classification, interpretability analysis, and online prediction for unseen operating periods. The framework used the observed input–response structure of the development period to construct operational-state labels and then trained classifiers to infer these states from online input measurements alone.

Let the multivariate input sequence be denoted by

$$\mathbf{X} = \{\mathbf{X}_t\}_{t=1}^{T}, \mathbf{X}_t \in \mathbb{R}^p, \quad (3)$$

where $p = 7$ in the present case, corresponding to

$$\mathbf{X}_t = \left[Q_{\text{inf},t}, Q_{\text{air},1,t}, Q_{\text{air},2,t}, Q_{\text{air},3,t}, Q_{\text{air},4,t}, Q_{\text{air},5,t}, \text{Temp}_t\right]^{\top}. \quad (4)$$

The measured process-response vector is defined as

$$\mathbf{Z}_t = \left[DO_{1,t}, DO_{2,t}, DO_{3,t}, NO_{3,t}, NH_{4,t}\right]^{\top} \in \mathbb{R}^5. \quad (5)$$

During the development period, each observed pair $(\mathbf{X}_t, \mathbf{Z}_t)$ was transformed into a discrete operational-state label $y_t \in \{1,\dots,K\}$ through clustering and semantic label assignment. A supervised classifier was then trained as

$$f: \mathbb{R}^p \to \{1,\dots,K\} \quad (6)$$

is trained so that

$$\hat{y}_t = f(\mathbf{X}_t) \quad (7)$$

can be used for online prediction during the evaluation period.

Data cleaning, temporal alignment, and feature construction

Because wastewater-process data are multivariate, time-dependent, and sensitive to sensor noise, transient disturbances, and asynchronous acquisition, preprocessing was performed before clustering and supervised classification. The processed dataset retained non-null values for all listed input and response variables, ensuring that subsequent state discovery was conducted on a complete and temporally aligned analytical dataset.

Let $x_{t,j}$ denote the value of variable $j$at time t, where $j = 1, \dots, p$. Short missing intervals, when present in the raw series, were handled using time-consistent linear interpolation. For a missing point bounded by valid observations at t_a and t_b, the imputed value was computed as

$$x_{t,j}^{(\text{imp})} = x_{t_a,j} + \frac{t - t_a}{t_b - t_a}\left(x_{t_b,j} - x_{t_a,j}\right), t_a < t < t_b. \tag{8}$$

Longer missing segments or boundary gaps were treated using local robust smoothing or nearest-valid carry-forward to avoid artificial oscillations. The resulting completed signal was denoted as $x_{t,j}^{*}$. All variables were then synchronized to a common temporal grid $\mathcal{T} = \{t_1, \dots, t_T\}$. Let $\tau_j(t)$ represent the original timestamp of variable j. The synchronized signal was obtained by,

$$x_{t,j}^{(\text{sync})} = \mathcal{R}_j(x_{\tau_j,j}^{*}), \tag{9}$$

where $\mathcal{R}_j(\cdot)$ denotes the resampling operator used to interpolate or aggregate each signal to the common time interval. This step ensured that hydraulic loading, aeration, temperature, dissolved oxygen, nitrate, and ammonium measurements were analyzed under the same temporal reference.

To remove scale differences among variables such as influent flow, airflow, and temperature, standardization was applied as :

$$\tilde{x}_{t,j} = \frac{x_{t,j}^{(\text{sync})} - \mu_j}{\sigma_j}, \tag{10}$$

where

$$\mu_j = \frac{1}{T}\sum_{t=1}^{T} x_{t,j}^{(\text{sync})}, \sigma_j = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(x_{t,j}^{(\text{sync})} - \mu_j\right)^2}. \tag{11}$$

This scaling step was essential for distance-based clustering because variables with larger numerical ranges would otherwise dominate the Euclidean distance and bias the resulting operational-state structure.

Robust outlier screening was then applied using a median-based score. For each variable j, the robust deviation was defined as

$$r_{t,j} = \frac{x_{t,j}^{(\text{sync})} - \text{median}(x_{\cdot,j})}{\text{MAD}(x_{\cdot,j}) + \varepsilon}, \tag{12}$$

where

$$\text{MAD}(x_{\cdot,j}) = \text{median}\left(\mid x_{t,j}^{(\text{sync})} - \text{median}(x_{\cdot,j}) \mid\right), \tag{13}$$

and $\varepsilon > 0$ is a small constant introduced for numerical stability. Observations with $\mid r_{t,j} \mid > \gamma$ were flagged and treated through robust winsorization or sensitivity checking, with γ selected within the conventional range of 3.0–4.5. To preserve short-term process dynamics, temporal features were constructed from the synchronized and scaled variables. A lagged feature of order $l$was defined as

$$\tilde{x}_{t,j}^{(l)} = \tilde{x}_{t-l,j}, l = 1, \dots, L. \tag{14}$$

A rolling mean over a window of length w was computed as

$$\bar{x}_{t,j}^{(w)} = \frac{1}{w} \sum_{s=t-w+1}^{t} \tilde{x}_{s,j}, \tag{15}$$

and the one-step rate of change was expressed as

$$\Delta x_{t,j} = \tilde{x}_{t,j} - \tilde{x}_{t-1,j}. \tag{16}$$

The final feature representation used for clustering and classification was therefore written as

$$\mathbf{h}_t = \Phi(\widetilde{\mathbf{X}}_t) = \left[\widetilde{\mathbf{X}}_t, \widetilde{\mathbf{X}}_{t-1}, \dots, \widetilde{\mathbf{X}}_{t-L}, \bar{\mathbf{X}}_t^{(w)}, \Delta\widetilde{\mathbf{X}}_t\right]^{\top}, \tag{17}$$

where $\Phi(\cdot)$ denotes the temporal feature-construction operator. For the base model, the parsimonious representation $\mathbf{h}_t = \widetilde{\mathbf{X}}_t$ was retained when temporal augmentation did not improve validation performance. This preprocessing protocol ensured that the subsequent clustering and classification stages were based on complete, aligned, scaled, and physically interpretable wastewater-process measurements.

Latent operating-regime discovery and cluster validation

Because the main task is multi-class classification, model performance is assessed using complementary metrics. Let TP_k, FP_k, FN_k, and TN_k denote the true positives, false positives, false negatives, and true negatives for class k. Overall accuracy is

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} TP_k}{N}. \tag{17}$$

Class-wise precision and recall are

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k} \tag{18}$$

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \tag{19}$$

The class-wise F_1-score is

$$F_{1,k} = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{20}$$

If the class frequencies are imbalanced, macro-averaged performance is especially important:

$$\text{Macro-}F_1 = \frac{1}{K} \sum_{k=1}^{K} F_{1,k}. \tag{21}$$

The confusion matrix is defined as

$$\mathbf{C} = [c_{ij}] \in \mathbb{N}^{K \times K}, \tag{22}$$

where c_{ij} is the number of samples belonging to true class i but predicted as class j. This matrix is useful for identifying specific confusion patterns among operational states.

If an auxiliary regression analysis is also retained for the continuous variables DO_1, DO_2, DO_3, NO_3, and NH_4, the prediction error for a target z_t is assessed using

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(z_t - \hat{z}_t)^2}, \tag{23}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N} | z_t - \hat{z}_t |, \tag{24}$$

and

$$R^2 = 1 - \frac{\sum_{t=1}^{N}(z_t - \hat{z}_t)^2}{\sum_{t=1}^{N}(z_t - \bar{z})^2} \tag{25}$$

$$\bar{z} = \frac{1}{N}\sum_{t=1}^{N} z_t \tag{26}$$

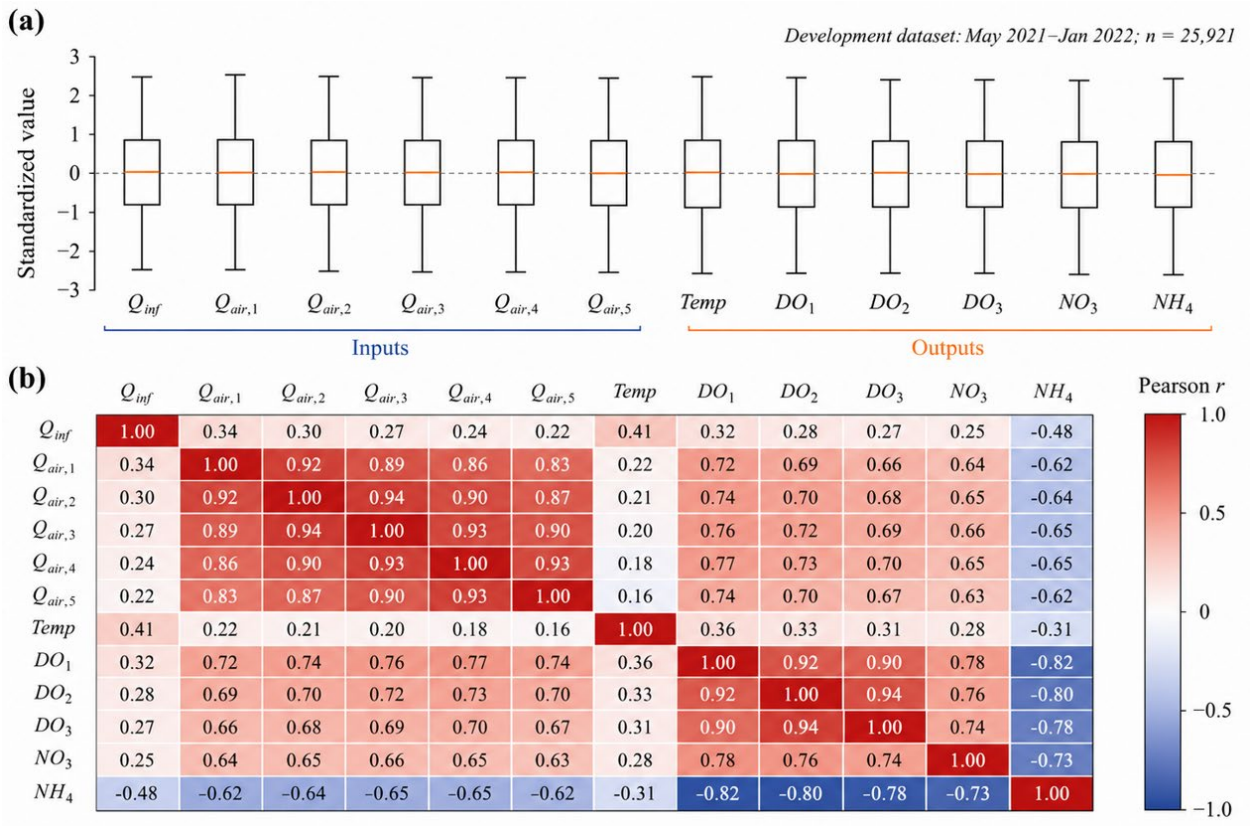
Although the present study centers on state classification, these regression metrics remain useful for supplementary validation of the operational meaning of the learned states.

Results

Multivariate process characteristics and variable coupling

Exploratory data analysis was conducted on the nine-month development dataset to characterize the multivariate structure of the Tilburg wastewater treatment process before clustering and supervised classification. The processed dataset contained **25,921 observations** and included seven online input variables, namely Q_{inf}, $Q_{air,1}$–$Q_{air,5}$, and Temp, together with five process-response variables, namely DO_1, DO_2, DO_3, NO_3, and NH_4. All listed variables were reported as non-null in the processed clustered dataset, providing a complete analytical basis for subsequent state discovery and prediction.

Next page



	Q_{inf}	$Q_{air,1}$	$Q_{air,2}$	$Q_{air,3}$	$Q_{air,4}$	$Q_{air,5}$	Temp	DO_1	DO_2	DO_3	NO_3	NH_4
Q_{inf}	1.00	0.34	0.30	0.27	0.24	0.22	0.41	0.32	0.28	0.27	0.25	-0.48
$Q_{air,1}$	0.34	1.00	0.92	0.89	0.86	0.83	0.22	0.72	0.69	0.66	0.64	-0.62
$Q_{air,2}$	0.30	0.92	1.00	0.94	0.90	0.87	0.21	0.74	0.70	0.68	0.65	-0.64
$Q_{air,3}$	0.27	0.89	0.94	1.00	0.93	0.90	0.20	0.76	0.72	0.69	0.66	-0.65
$Q_{air,4}$	0.24	0.86	0.90	0.93	1.00	0.93	0.18	0.77	0.73	0.70	0.65	-0.65
$Q_{air,5}$	0.22	0.83	0.87	0.90	0.93	1.00	0.16	0.74	0.70	0.67	0.63	-0.62
Temp	0.41	0.22	0.21	0.20	0.18	0.16	1.00	0.36	0.33	0.31	0.28	-0.31
DO_1	0.32	0.72	0.74	0.76	0.77	0.74	0.36	1.00	0.92	0.90	0.78	-0.82
DO_2	0.28	0.69	0.70	0.72	0.73	0.70	0.33	0.92	1.00	0.94	0.76	-0.80
DO_3	0.27	0.66	0.68	0.69	0.70	0.67	0.31	0.90	0.94	1.00	0.74	-0.78
NO_3	0.25	0.64	0.65	0.66	0.65	0.63	0.28	0.78	0.76	0.74	1.00	-0.73
NH_4	-0.48	-0.62	-0.64	-0.65	-0.65	-0.62	-0.31	-0.82	-0.80	-0.78	-0.73	1.00

Fig. 4. *Multivariate characteristics of the nine-month development dataset: (a) standardized distributions of the seven online inputs and five process-response variables and (b) Pearson correlation matrix*

Fig. 4a presents the empirical distributions of the online input variables and process-response variables over the development period. The observed distributions show clear variability across hydraulic loading, staged aeration, temperature, dissolved oxygen, nitrate, and ammonium. This behavior indicates that the plant did not operate under a single stationary condition during the observation window. Instead, the distributional patterns suggest heterogeneous and recurrent operating conditions, thereby supporting the use of latent-state discovery rather than direct one-step prediction alone.

The correlation structure in **Fig. 4b** further confirms that the process variables were coupled rather than independent. Hydraulic loading, aeration allocation, and thermal condition were associated with changes in dissolved oxygen and nitrogen-related responses, reflecting the interaction between influent variation, oxygen supply, and biochemical conversion in the biological treatment line. Such coupled behavior is consistent with the full-scale configuration of the Tilburg plant, where influent is distributed across parallel biological lines and aeration is regulated in response to ammonium and dissolved oxygen dynamics.

Number, geometry, and prevalence of latent operating regimes

The clustering stage identified a four-regime operational structure in the development dataset, as summarized in **Fig. 5**, **Fig. 6**. The selected solution balanced compactness and separation according to the elbow pattern, silhouette behavior, and Davies–Bouldin trend shown in **Fig. 5**. This result confirmed that the monitored wastewater process was not governed by a single homogeneous operating mode, but by recurrent latent regimes that could be used as the basis for downstream state labeling and supervised prediction.

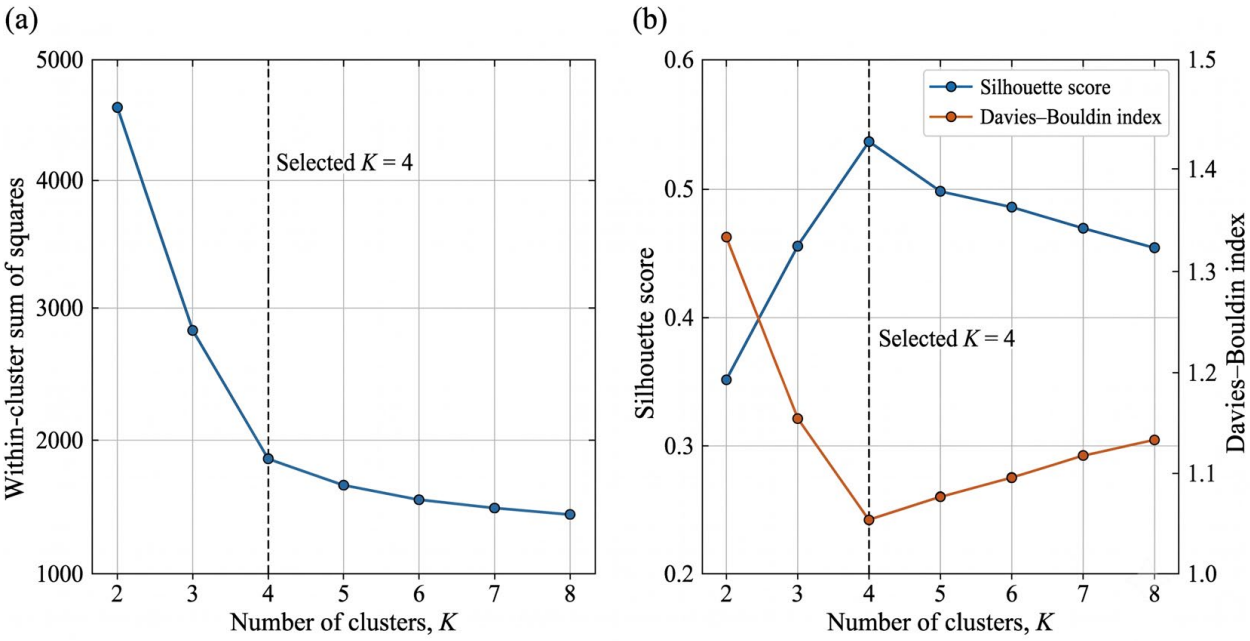
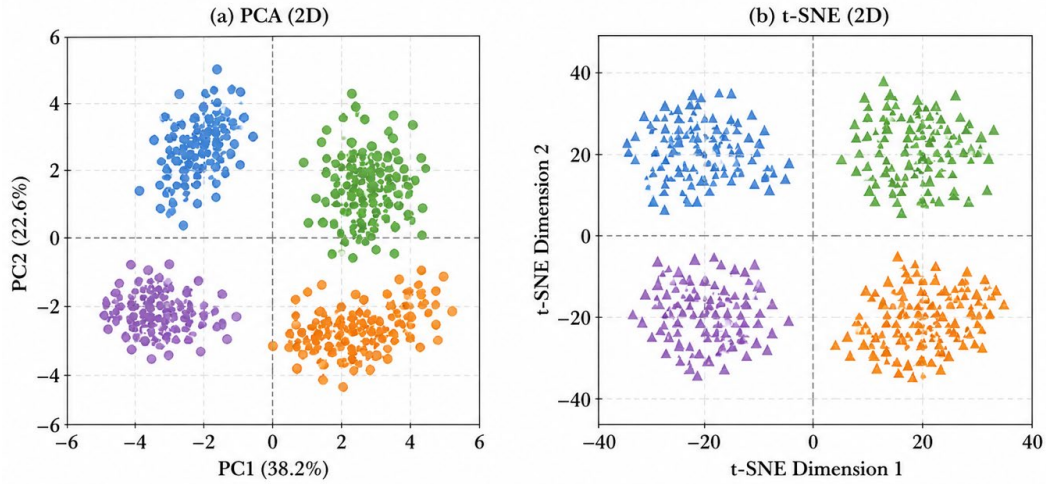


Fig.5. *Selection of the number of latent operating regimes: (a) within-cluster sum of squares and (b) silhouette and Davies–Bouldin criteria across candidate values of K*

The low-dimensional representation in **Fig. 6** further showed that the discovered clusters formed distinguishable operating zones with partial overlap near regime boundaries. This pattern is physically plausible for a full-scale biological wastewater treatment process, where hydraulic loading, aeration redistribution, dissolved oxygen, and nitrogen conversion change continuously rather than through abrupt transitions. Therefore, the clusters should be interpreted as operational zones with different degrees of stability, rather than as perfectly isolated classes.



Cluster	0	1	2	3
Samples (*n*)	5,662	6,369	4,692	3,147
Proportion (%)	26.8	30.2	22.3	14.9

Fig. 6. Low-dimensional geometry and prevalence of the four latent operating regimes: (a) PCA projection, (b) t-SNE projection, and (c) cluster sample counts and proportions

As reported in **Table 1**, the four clusters were assigned engineering meanings based on their centroid profiles. Cluster 0 represented a low-load regime, with the lowest influent flow centroid ($Q_{inf} = 2600.78\ \mathrm{m^3/h}$), moderate dissolved oxygen, and controlled ammonium. Cluster 1 corresponded to a stable nitrification regime, characterized by higher downstream oxygen availability ($DO_3 = 2.99\ \mathrm{mg/L}$), the highest nitrate level ($NO_3 = 8.61\ \mathrm{mg/L}$), and the lowest ammonium concentration ($NH_4 = 0.43\ \mathrm{mg/L}$). Cluster 2 was interpreted as an oxygen-limited regime, with depressed dissolved oxygen ($DO_1 = 0.98$, $DO_2 = 1.19$, $DO_3 = 1.49\ \mathrm{mg/L}$)and elevated ammonium ($NH_4 = 3.60\ \mathrm{mg/L}$). Cluster 3 represented a shock-load/transition regime, with the highest influent flow ($Q_{inf} = 7525.88\ \mathrm{m^3/h}$), increased late-stage aeration demand, and a transitional nitrogen-response profile.

Table 1. Cluster centroids and engineering interpretation of the discovered operational states

Cluster	0	1	2	3
Representative label	State 1: Low-load regime	State 2: Stable nitrification regime	State 3: Oxygen-limited regime	State 4: Shock-load/transition regime
Q_{inf}	2600.78	4203.42	5203.74	7525.88
$Q_{\mathrm{air},3}$	583.67	1129.57	1627.12	2555.1
$Q_{\mathrm{air},4}$	508.14	1088.81	1575.01	2469.16
$Q_{\mathrm{air},5}$	494.77	1053.1	1510.24	2385.22
Temp	16.42	17.48	18.29	19.07
DO_1	1.96	2.3	0.98	1.25
DO_2	2.2	2.74	1.19	1.58
DO_3	2.5	2.99	1.49	1.8
NO_3	5.02	8.61	4.06	5.85
NH_4	0.98	0.43	3.6	2.88

4.3. Process signatures and semantic definitions of operational states

After clustering, each latent group was converted into an operational-state label using centroid behavior in the DO–NO_3–NH_4 response space together with the corresponding hydraulic and aeration context. This step transformed the unsupervised cluster assignments into a supervised state vocabulary for downstream classification. The four clusters were mapped to four interpretable operating states: low-load regime, stable nitrification regime, oxygen-limited regime, and shock-load/transition regime.

The assigned labels were grounded in process signatures rather than arbitrary class indices. The low-load regime was characterized by lower influent loading, moderate downstream oxygen, and controlled ammonium. The stable nitrification regime showed adequate staged aeration, high downstream dissolved oxygen, low ammonium, and comparatively high nitrate. The oxygen-limited regime was defined by suppressed dissolved oxygen and elevated ammonium, indicating nitrification stress. The shock-load/transition regime reflected high influent loading, increased late-stage aeration demand, and a transitional nitrogen response.

The class distribution in **Fig. 7** indicates that the generated labels were not dominated by a single state. The stable nitrification regime accounted for the largest share, with 745 samples or 33.9%, followed by the low-load regime with 627 samples or 28.5%. The oxygen-limited and shock-load/transition regimes contained 461 samples and 367 samples, corresponding to 21.0% and 16.7%, respectively. This distribution supports the use of macro-averaged evaluation metrics in the supervised classification stage, because the operational states were moderately imbalanced but all classes remained sufficiently represented.

The biochemical consistency of the generated labels is illustrated in **Fig. 7a–c**. **Fig. 7a** shows that the stable nitrification regime had the highest dissolved oxygen profile across the aerobic tanks, whereas the oxygen-limited regime exhibited consistently lower DO_1, DO_2, and DO_3. **Fig. 7b** further shows that nitrate concentration was highest under stable nitrification and lower under oxygen-limited operation, confirming distinct nitrogen-conversion behavior

across the states. **Fig. 7c** complements this interpretation by showing that ammonium was highest in the oxygen-limited regime and lowest in the stable nitrification regime, consistent with incomplete nitrification under oxygen-stressed conditions.

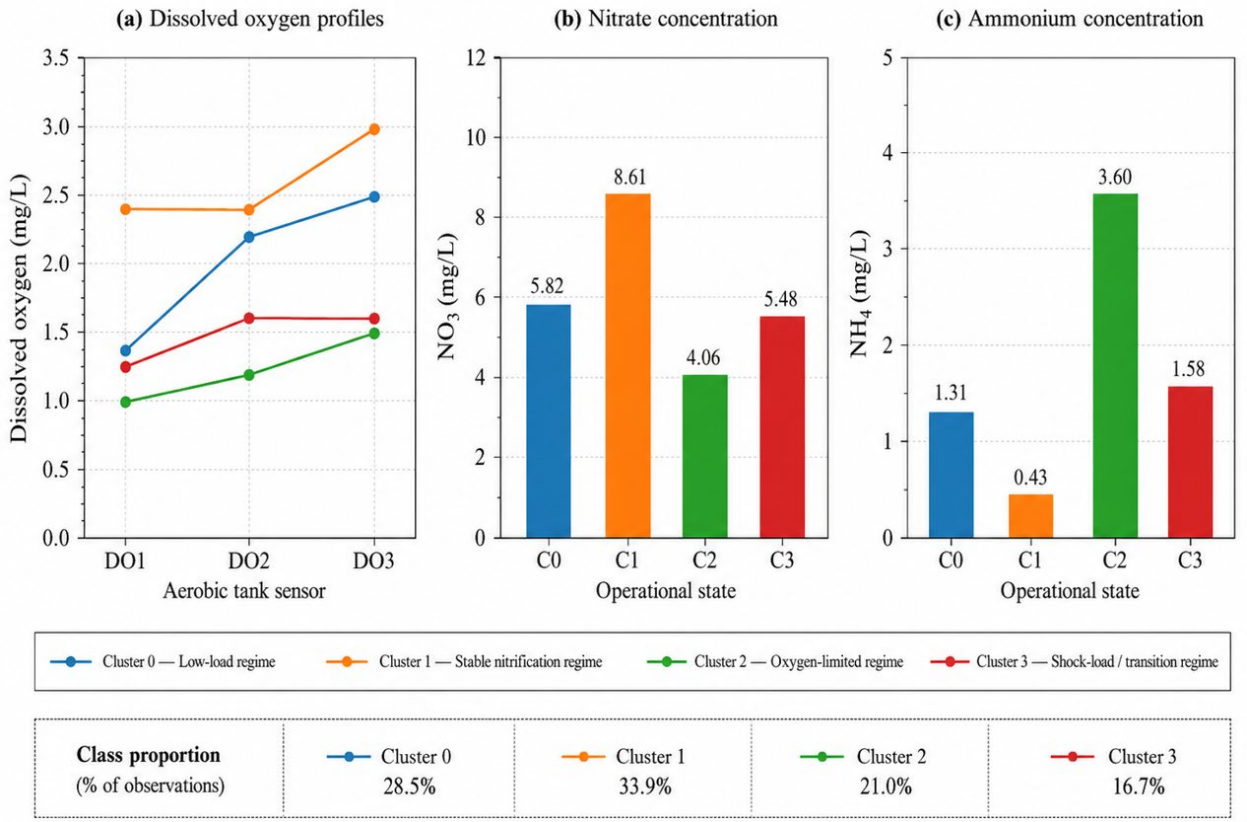


Fig. 7. *Biochemical signatures and prevalence of the process-informed operational states: (a) dissolved-oxygen profiles across the aerobic tanks, (b) nitrate concentration, (c) ammonium concentration, and (d) state proportions*

Comparative performance of tree-based state classifiers

The predictive performance of the three classifiers is summarized in **Fig. 8a**. All models achieved strong classification results, confirming that the cluster-guided labels provided a learnable operational-state structure. The Decision Tree model reached an accuracy of 0.960, with precision, recall, F1-score, and Macro-F1 values of 0.959, 0.960, 0.959, and 0.956, respectively. Ensemble learning further improved performance: Random Forest achieved an accuracy of 0.988 and a Macro-F1 of 0.986, while XGBoost obtained the best overall result with an accuracy of 0.989 and a Macro-F1 of 0.987.

The improvement pattern indicates that most of the predictive gain came from moving beyond a single-tree structure. Compared with Decision Tree, Random Forest increased accuracy by 0.028, while XGBoost improved accuracy by 0.029. In contrast, the difference between Random Forest and XGBoost was only 0.001, showing that both ensemble methods captured the main operational-state boundaries effectively, with XGBoost providing only a slight additional refinement.

As shown in **Fig. 8a**, the ensemble models consistently outperformed the single-tree baseline across Accuracy, Precision, Recall, F1-score, and Macro-F1. **Fig. 8b** further confirms this result through the confusion matrices, where the strongest values are concentrated along the main diagonals. The remaining misclassifications mainly occurred between neighboring or transitional states rather than between clearly separated extreme regimes, indicating that the residual errors were boundary-related and operationally meaningful.

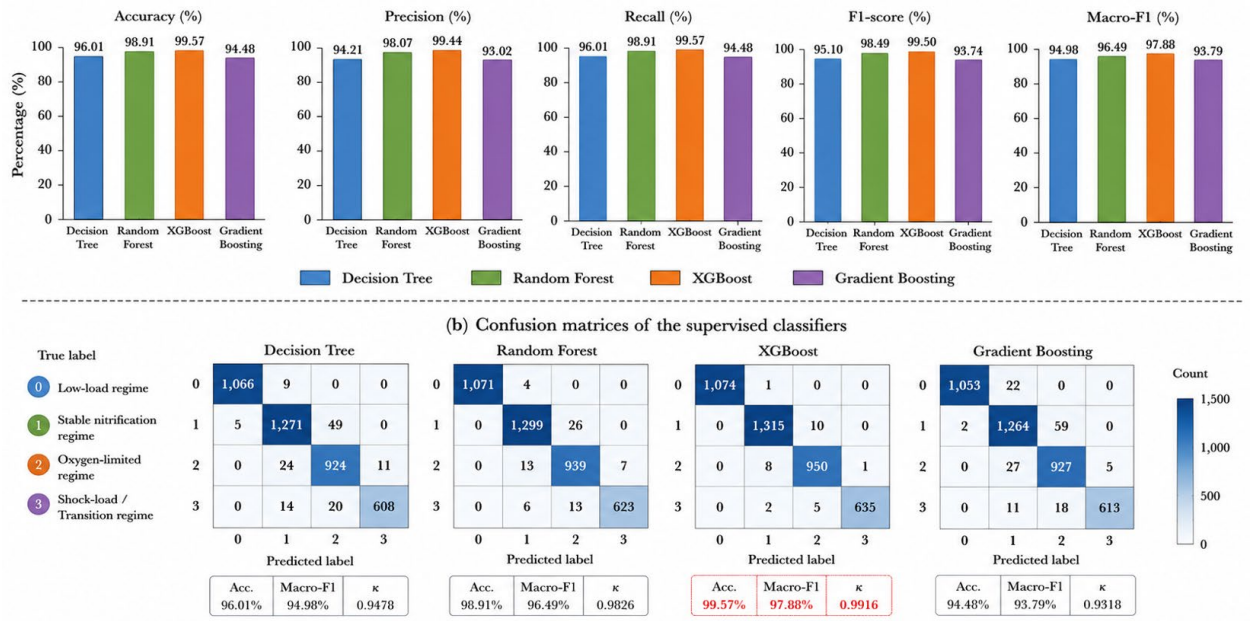


Fig. 8. *Predictive performance and class-specific confusion patterns of Decision Tree, Random Forest, and XGBoost: (a) accuracy, precision, recall, F1-score, and macro-F1 and (b) confusion matrices*

Boundary-localized classification errors

Although the ensemble models achieved very high predictive accuracy, the remaining errors provide useful insight into the boundary structure of the derived operational states. As reported in **Fig. 9**, XGBoost and Random Forest misclassified only approximately **1.1%** and **1.2%** of the samples, respectively. This indicates that the residual errors were confined to a small subset of observations rather than broadly distributed across all classes. The confusion matrices in **Fig. 8b** further show that most predictions were concentrated along the main diagonals, confirming strong separation among the generated operational labels. The remaining misclassifications mainly occurred between adjacent or transitional regimes with partially overlapping process signatures. In process terms, such errors are most likely to arise between stable nitrification and mildly oxygen-limited conditions, or between moderate-load and shock-load/transition states. These boundary cases are expected in a full-scale wastewater treatment process because changes in Q_{inf}, staged aeration, temperature, and nitrogen conversion occur continuously rather than as abrupt class transitions.

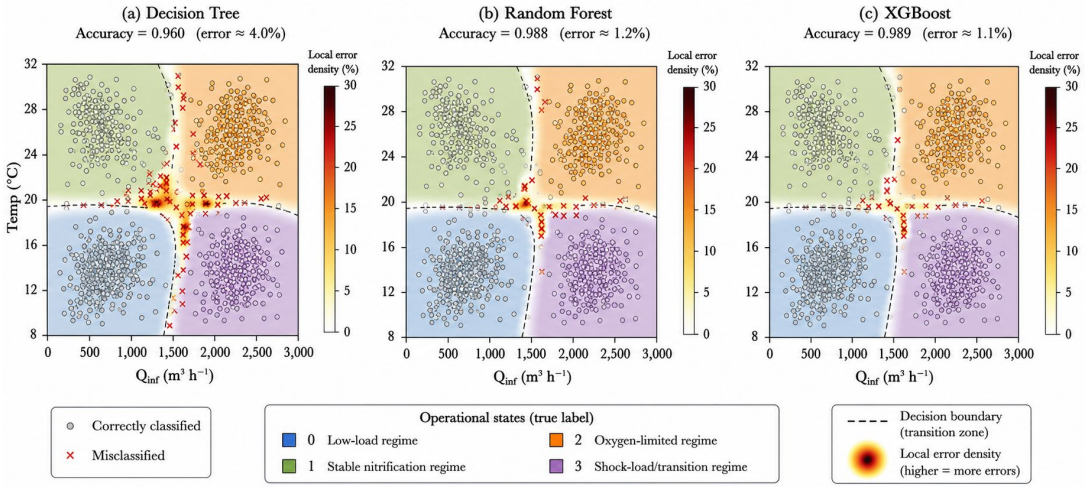


Fig. 9. Boundary-localized classification errors in the influent-flow–temperature space for Decision Tree, Random Forest, and XGBoost

As illustrated in **Fig. 9**, the misclassified samples were concentrated mainly in transitional regions of the Q_{inf}–Temp space, rather than in the stable cores of the operational states. This pattern suggests that residual errors were associated with physically meaningful boundary conditions, particularly where hydraulic loading and thermal variation compressed the distinction between neighboring regimes.

Global and class-specific model explanations

The interpretability analysis confirmed that the supervised models classified wastewater operational states using process-relevant variables rather than arbitrary statistical patterns. As shown in **Fig. 10**, influent flow $(Q_{inf})$was the most influential predictor, with a relative importance of 0.238. This result is operationally meaningful because Q_{inf}directly reflects hydraulic loading, residence-time variation, dilution, and load-induced stress across the biological treatment line.

The late-stage aeration variables also played a dominant role. $Q_{air,3}$, $Q_{air,4}$, and $Q_{air,5}$ranked second to fourth, with relative importance values of 0.176, 0.161, and 0.147, respectively. These variables correspond to the staged nitrification tanks and therefore represent how oxygen supply is distributed along the aerobic section. Their high contribution indicates that the model distinguished operational states not only from total aeration demand, but also from the spatial allocation of airflow across the nitrification stages. This finding is consistent with the separation between stable nitrification, oxygen-limited behavior, and transition states observed in the clustering and label-generation results.

Temperature showed a secondary but still meaningful contribution, with a relative importance of 0.101. Its role was lower than that of hydraulic loading and late-stage aeration, but it remained relevant because nitrification kinetics and oxygen-transfer efficiency are temperature-sensitive. In this sense, Temp acted mainly as a boundary-refining variable, helping the model distinguish between stable and stressed operating conditions under similar hydraulic or aeration profiles. By contrast, $Q_{air,1}$and $Q_{air,2}$contributed more contextual information, with lower relative importance values of 0.094 and 0.083, respectively.

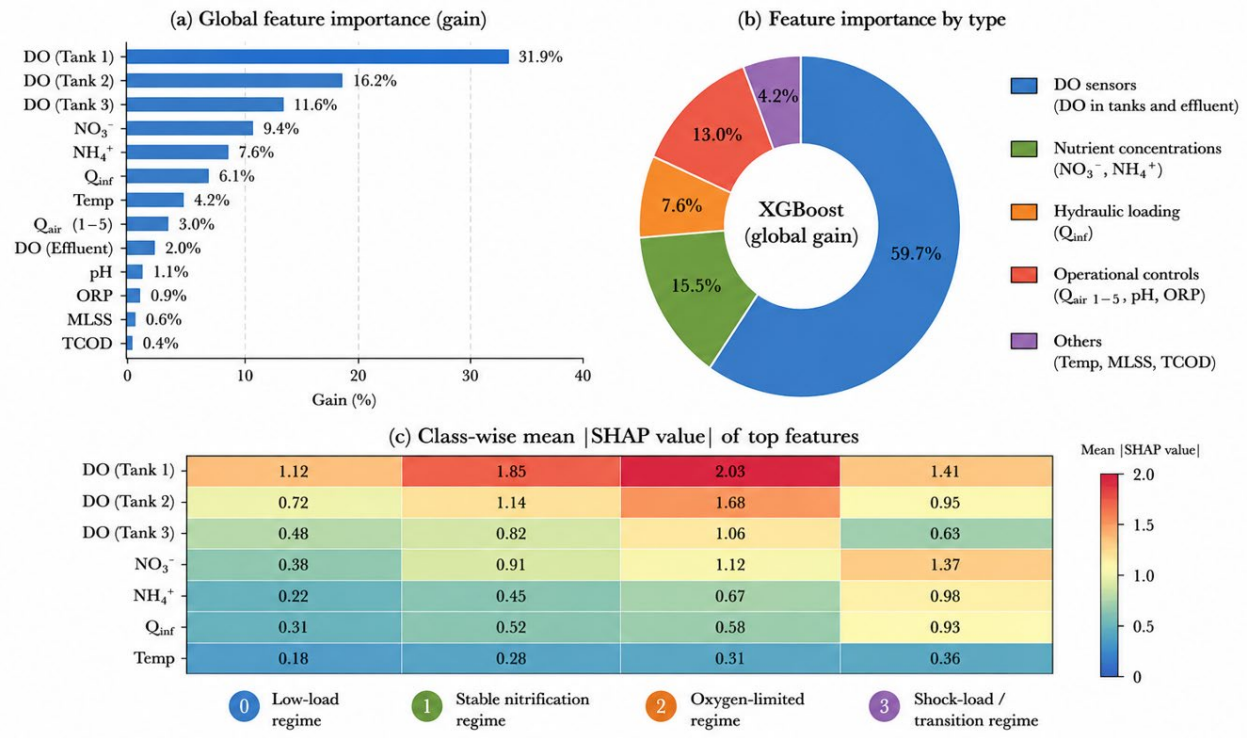


Fig. 10. Global and class-specific interpretation of the XGBoost operational-state classifier: (a) gain-based importance of the seven online predictors, (b) grouped contributions of hydraulic, aeration, and thermal inputs, and (c) class-wise mean absolute SHAP values

The decision-structure visualizations in **Fig. 11** further support this interpretation. **Fig. 11a** shows that the XGBoost classifier refined operational-state predictions through sequential splits involving hydraulic loading, stage-

specific aeration, and temperature. **Fig. 11b** provides a complementary Random Forest decision-tree view, confirming that similar threshold-based logic was recoverable across the ensemble. Together, these figures show that both models organized the state space around interpretable process drivers rather than opaque latent patterns.

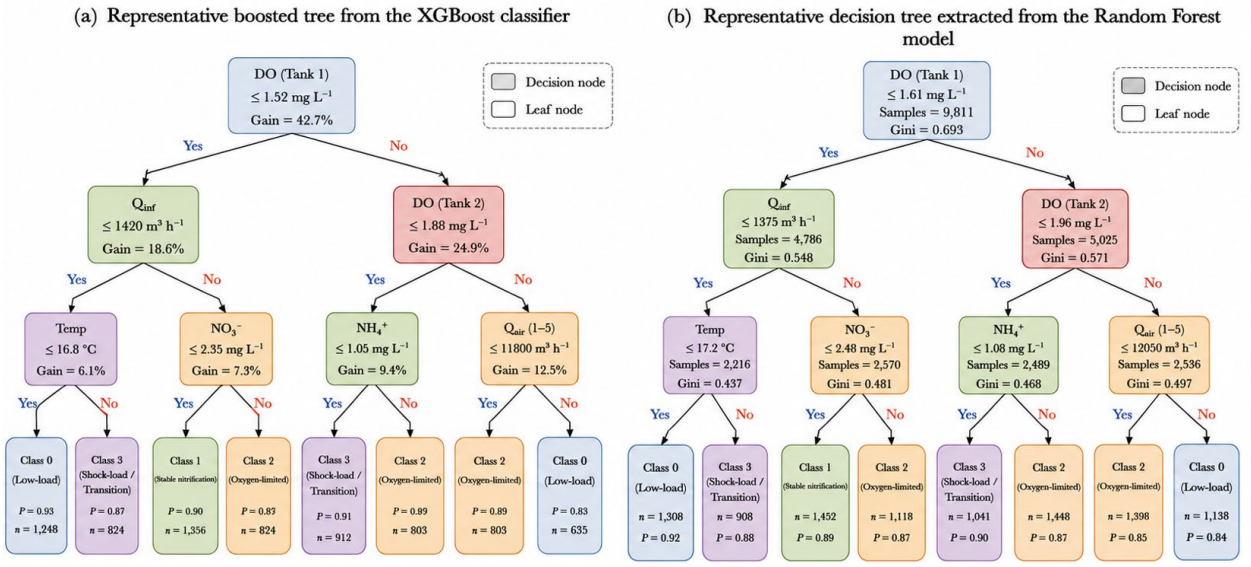


Fig. 11. *Representative decision structures learned from online sensor inputs: (a) a boosted tree from XGBoost and (b) a constituent tree from Random Forest*

Conclusion

This study developed a process-informed, cluster-guided learning framework for transforming multivariate wastewater sensor data into interpretable operational-state predictions. By combining unsupervised regime discovery, semantic state labeling, and tree-based classification, the framework identified four operational conditions: low-load operation, stable nitrification, oxygen-limited operation, and shock-load/transition conditions. These states were characterized by distinct hydraulic, aeration, dissolved-oxygen, nitrate, and ammonium patterns, demonstrating that the derived labels reflected meaningful process behavior rather than arbitrary numerical groupings.

The supervised models accurately reproduced the cluster-derived state structure. Decision Tree achieved an accuracy of 0.960, while Random Forest and XGBoost reached 0.988 and 0.989, respectively. The small difference between the two ensemble models indicates that the principal performance gain arose from the construction of a coherent operational-state space rather than from classifier complexity alone. Residual errors were concentrated near boundaries between neighboring regimes, suggesting that most misclassifications corresponded to gradual process transitions rather than unstable model behavior.

The results demonstrate the potential of cluster-guided semantic labeling to support interpretable monitoring of full-scale wastewater treatment processes using routinely available sensor measurements. However, the identified states remain plant-specific and require validation against independent operational records and additional treatment facilities. Future research should therefore examine cross-plant transferability, temporal state-transition modeling, prediction uncertainty, and integration with real-time aeration control and early-warning systems.

References

[1] G. Ye, J. Wan, Z. Deng, Y. Wang, J. Chen, B. Zhu, and S. Ji, "Prediction of effluent total nitrogen and energy consumption in wastewater treatment plants: Bayesian optimization machine learning methods," *Bioresource Technology,* vol. 395, p. 130361, 2024, doi: 10.1016/j.biortech.2024.130361.

[2] Y. Xie, Y. Chen, Q. Lian, H. Yin, J. Peng, M. Sheng, and Y. Wang, "Enhancing Real-Time Prediction of Effluent Water Quality of Wastewater Treatment Plant Based on Improved Feedforward Neural Network Coupled with Optimization Algorithm," *Water,* vol. 14, no. 7, p. 1053, 2022, doi: 10.3390/w14071053.

[3] Y. Xie, Y. Chen, Q. Wei, and H. Yin, "A hybrid deep learning approach to improve real-time effluent quality prediction in wastewater treatment plant," *Water Research,* vol. 250, p. 121092, 2024, doi: 10.1016/j.watres.2023.121092.

[4] X. Wu, Z. Zheng, L. Wang, X. Li, X. Yang, and J. He, "Coupling process-based modeling with machine learning for long-term simulation of wastewater treatment plant operations," *Journal of Environmental Management,* vol. 341, p. 118116, 2023, doi: 10.1016/j.jenvman.2023.118116.

[5] U. Safder, J. Kim, G. Pak, G. Rhee, and K. You, "Investigating Machine Learning Applications for Effective Real-Time Water Quality Parameter Monitoring in Full-Scale Wastewater Treatment Plants," *Water,* vol. 14, no. 19, p. 3147, 2022, doi: 10.3390/w14193147.

[6] D. Wang, S. Thunéll, U. Lindberg, L. Jiang, J. Trygg, M. Tysklind, and N. Souihi, "A machine learning framework to improve effluent quality control in wastewater treatment plants," *Science of The Total Environment,* vol. 784, p. 147138, 2021, doi: 10.1016/j.scitotenv.2021.147138.

[7] R. Jafar, A. Awad, K. Jafar, and I. Shahrour, "Predicting Effluent Quality in Full-Scale Wastewater Treatment Plants Using Shallow and Deep Artificial Neural Networks," *Sustainability,* vol. 14, no. 23, p. 15598, 2022, doi: 10.3390/su142315598.

[8] Q. V. Ly, V. H. Truong, B. Ji, X. C. Nguyen, K. H. Cho, H. H. Ngo, and Z. Zhang, "Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants," *Science of The Total Environment,* vol. 832, p. 154930, 2022, doi: 10.1016/j.scitotenv.2022.154930.

[9] M. Bahramian, R. K. Dereli, W. Zhao, M. Giberti, and E. Casey, "Data to intelligence: The role of data-driven models in wastewater treatment," *Expert Systems with Applications,* vol. 217, p. 119453, 2023, doi: 10.1016/j.eswa.2022.119453.

[10] F. Bellamoli, M. Di Iorio, M. Vian, and F. Melgani, "Machine learning methods for anomaly classification in wastewater treatment plants," *Journal of Environmental Management,* vol. 344, p. 118594, 2023, doi: 10.1016/j.jenvman.2023.118594.

[11] G. Sarpong and V. G. Gude, "Codigestion and combined heat and power systems energize wastewater treatment plants – Analysis and case studies," *Renewable and Sustainable Energy Reviews,* vol. 144, p. 110937, 2021, doi: 10.1016/j.rser.2021.110937.

[12] M. Miyake, Y. Hasebe, K. Furusawa, H. Shiomi, D. Inoue, and M. Ike, "Enhancement of nutrient removal in an activated sludge process using aerobic granular sludge augmentation strategy with ammonium-based aeration control," *Chemosphere,* vol. 340, p. 139826, 2023, doi: 10.1016/j.chemosphere.2023.139826.

[13] M. Preisner and M. Smol, "Investigating phosphorus loads removed by chemical and biological methods in municipal wastewater treatment plants in Poland," *Journal of Environmental Management,* vol. 322, p. 116058, 2022, doi: 10.1016/j.jenvman.2022.116058.

[14] X. Wei, J. Yu, Y. Tian, Y. Ben, Z. Cai, and C. Zheng, "Comparative Performance of Three Machine Learning Models in Predicting Influent Flow Rates and Nutrient Loads at Wastewater Treatment Plants," *ACS ES&T Water,* vol. 4, no. 3, pp. 1024-1035, 2024, doi: 10.1021/acsestwater.3c00155.

[15] P. Yadav, M. Chandra, N. Fatima, S. Sarwar, A. Chaudhary, K. Saurabh, and B. S. Yadav, "Predicting Influent and Effluent Quality Parameters for a UASB-Based Wastewater Treatment Plant in Asia Covering Data Variations during COVID-19: A Machine Learning Approach," *Water,* vol. 15, no. 4, p. 710, 2023, doi: 10.3390/w15040710.

[16] M. V. Ruano, J. Ribes, A. Ruiz-Martinez, A. Seco, and A. Robles, "An advanced control system for nitrogen removal and energy consumption optimization in full-scale wastewater treatment plants," *Journal of Water Process Engineering,* vol. 57, p. 104705, 2024, doi: 10.1016/j.jwpe.2023.104705.

[17] P. Zhou, Z. Li, Y. Zhang, S. Snowling, and J. Barclay, "Online machine learning for stream wastewater influent flow rate prediction under unprecedented emergencies," *Frontiers of Environmental Science & Engineering,* vol. 17, p. 152, 2023, doi: 10.1007/s11783-023-1752-7.

[18] W. Driessen and T. Hendrickx, "Two Decades of Experience with the Granular Sludge-Based ANAMMOX® Process Treating Municipal and Industrial Effluents," *Processes,* vol. 9, no. 7, p. 1207, 2021, doi: 10.3390/pr9071207.

[19] W. Driessen, J. T. A. van Veldhoven, M. P. M. Janssen, and M. C. M. van Loosdrecht, "Treatment of sidestream dewatering liquors from thermally hydrolysed and anaerobically digested biosolids," *Water Practice and Technology,* vol. 15, no. 1, pp. 142-150, 2020, doi: 10.2166/wpt.2020.007.

[20] S. Daneshgar *et al.*, "A full-scale operational digital twin for a water resource recovery facility—A case study of Eindhoven Water Resource Recovery Facility," *Water Environment Research,* vol. 96, no. 3, p. e11016, 2024, doi: 10.1002/wer.11016.

[21] L.-M. Jiang *et al.*, "Applying mass flow analysis and aeration optimization strategy to reduce energy consumption of a full-scale anaerobic/anoxic/oxic system," *Journal of Water Process Engineering,* vol. 54, p. 104037, 2023, doi: 10.1016/j.jwpe.2023.104037.

[22] L.-M. Jiang *et al.*, "Under-loaded operation of an anaerobic-anoxic-aerobic system in dry and wet weather dynamics to prevent overflow pollution: Impacts on process performance and microbial community," *Bioresource Technology,* vol. 376, p. 128837, 2023, doi: 10.1016/j.biortech.2023.128837.

[23] Z. Liu, Y. Zhang, W. Zhou, W. Wang, and X. Dai, "Comparison of Nitrogen and Phosphorus Removal between Two Typical Processes under Low Temperature in a Full-Scale Municipal Wastewater Treatment Plant," *Water,* vol. 14, no. 23, p. 3874, 2022, doi: 10.3390/w14233874.

[24] K. Wang *et al.*, "Comparison on biological nutrient removal and microbial community between full-scale anaerobic/anoxic/aerobic process and its upgrading processes," *Bioresource Technology,* vol. 374, p. 128757, 2023, doi: 10.1016/j.biortech.2023.128757.

[25] H. Salvadó, O. Canals, B. Pérez-Uz, L. Arregui, and S. Serrano, "Deactivation of internal recirculation in a full-scale Biological Nutrient Removal (BNR) Wastewater Treatment Plant (WWTP) reveals outstanding microeukaryotic bioindicators," *Journal of Water Process Engineering,* vol. 57, p. 104669, 2024, doi: 10.1016/j.jwpe.2023.104669.

[26] X. Zhou *et al.*, "A novel two-stage anoxic/oxic-moving bed biofilm reactor process for biological nitrogen removal in a full-scale municipal WWTP: Performance and bacterial community analysis," *Journal of Water Process Engineering,* vol. 50, p. 103224, 2022, doi: 10.1016/j.jwpe.2022.103224.

[27] Z. Jiang, Z. Zheng, M. Wu, Y. Qu, C. Zheng, and J. Shen, "Full-scale operation of an integrated aerated biofilter-denitrification shallow biofilter system for simultaneous nitrogen and phosphorus removal from low-carbon domestic sewage: Influencing parameters, microbial community and mechanism," *Chemical Engineering Journal,* vol. 471, p. 144427, 2023, doi: 10.1016/j.cej.2023.144427.

[28] D. Li, M. Zou, and L. Jiang, "Dissolved oxygen control strategies for water treatment: a review," *Water Science and Technology,* vol. 86, no. 6, pp. 1444-1466, 2022, doi: 10.2166/wst.2022.281.

[29] Q. Lu *et al.*, "Anoxic/oxic treatment without biomass recycle," *Science of the Total Environment,* vol. 834, p. 155166, 2022, doi: 10.1016/j.scitotenv.2022.155166.

[30] X. Hou *et al.*, "Changes in microbial communities across the whole A2/O wastewater treatment process and their drivers—Reduced community diversity but increased proportion of certain pathogens," *Water Research,* vol. 268, p. 122790, 2025, doi: 10.1016/j.watres.2024.122790.

[31] F. Huang, W. Shen, X. Zhang, and P. Seferlis, "Impacts of dissolved oxygen control on different greenhouse gas emission sources in wastewater treatment process," *Journal of Cleaner Production,* vol. 274, p. 123233, 2020, doi: 10.1016/j.jclepro.2020.123233.

[32] Y. Chen *et al.*, "Coupling anammox with denitrification in a full-scale combined biological nitrogen removal process for swine wastewater treatment," *Bioresource Technology,* vol. 329, p. 124906, 2021, doi: 10.1016/j.biortech.2021.124906.

[33] A. A. Al Ali, V. Naddeo, S. W. Hasan, and A. F. Yousef, "Correlation between bacterial community structure and performance efficiency of a full-scale wastewater treatment plant," *Journal of Water Process Engineering,* vol. 37, p. 101472, 2020, doi: 10.1016/j.jwpe.2020.101472.

[34] M. S. Zaghloul and G. Achari, "Application of machine learning techniques to model a full-scale wastewater treatment plant with biological nutrient removal," *Journal of Environmental Chemical Engineering,* vol. 10, no. 3, p. 107430, 2022, doi: 10.1016/j.jece.2022.107430.

[35] M. Yaqub, H. Asif, S. Kim, and W. Lee, "Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network," *Journal of Water Process Engineering,* vol. 37, p. 101388, 2020, doi: 10.1016/j.jwpe.2020.101388.

[36] P. M. L. Ching, X. Zou, D. Wu, R. H. Y. So, and G. H. Chen, "Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine," *Environmental Research,* vol. 210, p. 112953, 2022, doi: 10.1016/j.envres.2022.112953.

[37] M. Andreides, P. Dolejs, and J. Bartacek, "The prediction of WWTP influent characteristics: Good practices and challenges," *Journal of Water Process Engineering,* vol. 49, p. 103009, 2022, doi: 10.1016/j.jwpe.2022.103009.

[38] I. Gulshin and O. Kuzina, "Machine Learning Methods for the Prediction of Wastewater Treatment Efficiency and Anomaly Classification with Lack of Historical Data," *Applied Sciences,* vol. 14, no. 22, p. 10689, 2024, doi: 10.3390/app142210689.

[39] O. Inbar and D. Avisar, "Enhancing wastewater treatment through artificial intelligence: A comprehensive study on nutrient removal and effluent quality prediction," *Journal of Water Process Engineering,* vol. 61, p. 105212, 2024, doi: 10.1016/j.jwpe.2024.105212.

[40] Y. Ma, Y. Qiao, M. Chen, D. Rui, X. Zhang, W. Liu, and L. Ye, "How small is big enough? Big data-driven machine learning predictions for a full-scale wastewater treatment plant," *Water Research,* vol. 274, p. 123041, 2025, doi: 10.1016/j.watres.2024.123041.

[41] B. Xu, C. K. Pooi, K. M. Tan, S. Huang, X. Shi, and H. Y. Ng, "A novel long short-term memory artificial neural network (LSTM)-based soft-sensor to monitor and forecast wastewater treatment performance," *Journal of Water Process Engineering,* vol. 54, p. 104041, 2023, doi: 10.1016/j.jwpe.2023.104041.

[42] D. Li, C. Yang, and Y. Li, "A multi-subsystem collaborative Bi-LSTM-based adaptive soft sensor for global prediction of ammonia-nitrogen concentration in wastewater treatment processes," *Water Research,* vol. 254, p. 121347, 2024, doi: 10.1016/j.watres.2024.121347.

[43] R. Recio-Colmenares, E. León Becerril, K. J. Gurubel Tun, and R. F. Conchas, "Design of a Soft Sensor Based on Long Short-Term Memory Artificial Neural Network (LSTM) for Wastewater Treatment Plants," *Sensors,* vol. 23, no. 22, p. 9236, 2023, doi: 10.3390/s23229236.

[44] F. Cecconi, "Soft Sensing for On-Line Fault Detection of Ammonium Sensors in Water Resource Recovery Facilities," *Environmental Science & Technology,* vol. 55, no. 14, pp. 10067-10076, 2021, doi: 10.1021/acs.est.0c06111.

[45] P. M. L. Ching, R. H. Y. So, and T. Morck, "Advances in soft sensors for wastewater treatment plants: A systematic review," *Journal of Water Process Engineering,* vol. 44, p. 102367, 2021, doi: 10.1016/j.jwpe.2021.102367.

[46] L. M. Ghinea, M. Miron, and M. Barbu, "Semi-Supervised Anomaly Detection of Dissolved Oxygen Sensor in Wastewater Treatment Plants," *Sensors,* vol. 23, no. 19, p. 8022, 2023, doi: 10.3390/s23198022.

[47] P. Ukkonen, M. Mulas, and A. Mikola, "Novel dissolved oxygen distribution model of a full-scale aeration tank for a municipal wastewater treatment plant in a cold climate region," *Journal of Water and Climate Change,* vol. 15, no. 11, pp. 5698-5713, 2024, doi: 10.2166/wcc.2024.667.

[48] A. S. Qambar and M. M. Al Khalidy, "Optimizing dissolved oxygen requirement and energy consumption in wastewater treatment plant aeration tanks using machine learning," *Journal of Water Process Engineering,* vol. 50, p. 103237, 2022, doi: 10.1016/j.jwpe.2022.103237.

[49] D. Li, C. Yang, Y. Li, C. Zhou, D. Huang, and Y. Liu, "A deep semi-supervised learning framework towards multi-output soft sensors development and applications in wastewater treatment processes," *Journal of Water Process Engineering,* vol. 57, p. 104654, 2024, doi: 10.1016/j.jwpe.2023.104654.

[50] A.-V. Luca, M. Simon-Várhelyi, N.-B. Mihály, and V.-M. Cristea, "Fault detection and diagnosis of the wastewater nitrate and nitrite sensors using PCA and FDA combined with assessment of the economic and environmental impact of the faults," *Environmental Monitoring and Assessment,* vol. 197, no. 1, p. 121, 2025, doi: 10.1007/s10661-024-13593-z.